Régression ou corrélation

Professeur E. Albuisson

> CHU et Faculté de Médecine

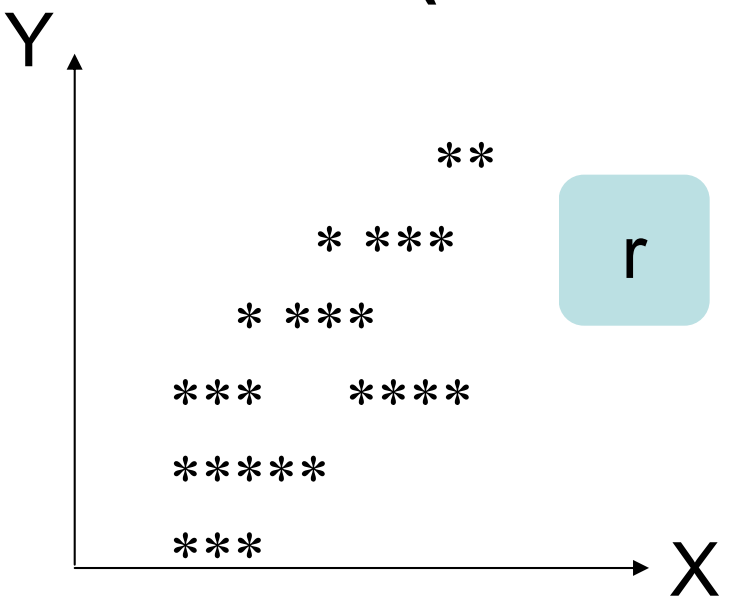
Nancy-Université
Université Henri Poincaré

ministère éducation nationale enseignement supérieur recherche

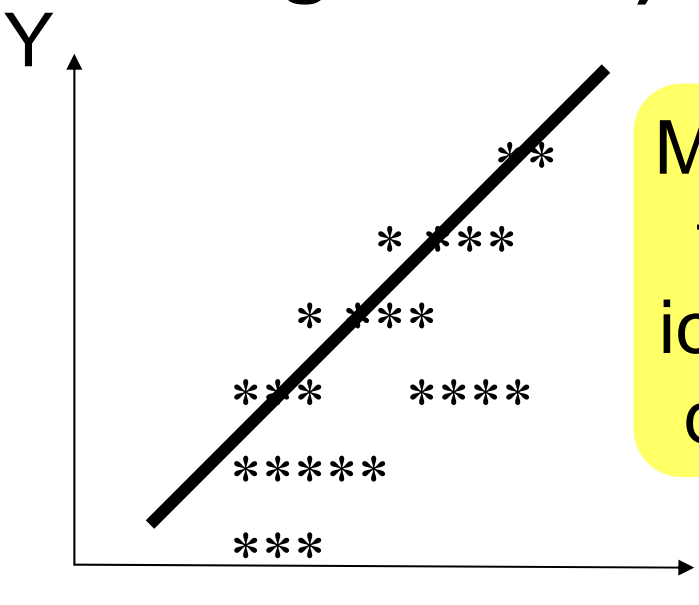
Principe général pour ces deux approches (corrélation ou régression)

Rechercher l'existence	Observation ou expérience
d'une liaison (relation, dépendance)	A définir
entre deux variables	On dit alors 'simple'
quantitatives	Nature des variables
X et Y	Aléatoires(ou non)Rôles(idem ou non)
appariées	n couples (x_i, y_i) de (X , Y)
ayant ou non la même unité	Parmi les rares approches à le permettre

Deux approches différentes (corrélation ou régression)



r



Modèle f (X) ici une droite

Nuage de points: X et Y sont interchangeables (rôles identiques). Calcul du coefficient de corrélation linéaire r

Nuage de points : X et Y ne sont pas interchangeables (rôles différents). Calcul des paramètres du modèle f(X). X peut être contrôlée

Régression: La connaissance de la valeur prise par X permet-elle de prédire la valeur prise par Y ?

$$\hat{Y} = f(X)$$

Y est la variable 'à expliquer' ou critère. $\hat{Y}$ est la prévision de Y par X en utilisant le modèle

X est la variable 'explicative' ou prédicteur

MODELE : RECHERCHE DE f( ) LA PLUS APPROPRIEE.
LINEAIRE, LOGARITHMIQUE, INVERSE,
CUBIQUE, PUISSANCE, LOGISTIQUE, EXPONENTIELLE,…

X aléatoire ou contrôlée?

En toute rigueur, les n couples (xi, yi) constituent un échantillon d'observations qui sont des réalisations de (X, Y), X et Y étant des variables aléatoires. Il est important de noter que *la corrélation ne s'appliquera que dans ce cas*.

Si la variable X est contrôlée par l'expérimentateur:

ex: dose croissante de médicament: d1, …, dk

ex: temps: t1, … , tk

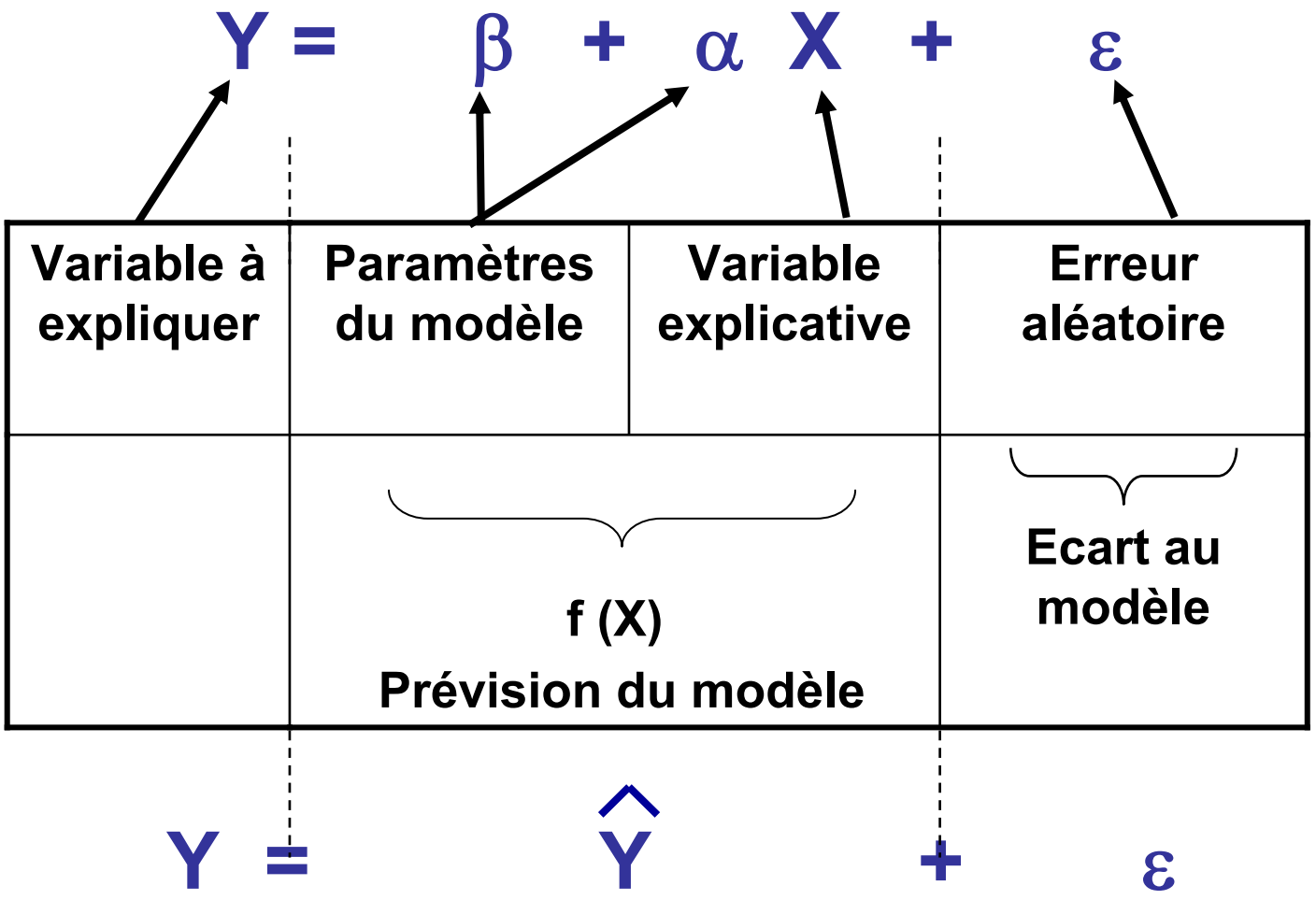
alors X n'est pas aléatoire et il s'agit plus d'un *modèle linéaire* que d'une *régression linéaire.*

Remarque: La méthode des moindres carrés utilisée pour rechercher les paramètres du modèle s'applique aussi bien au *modèle linéaire* qu'à la *régression linéaire*.

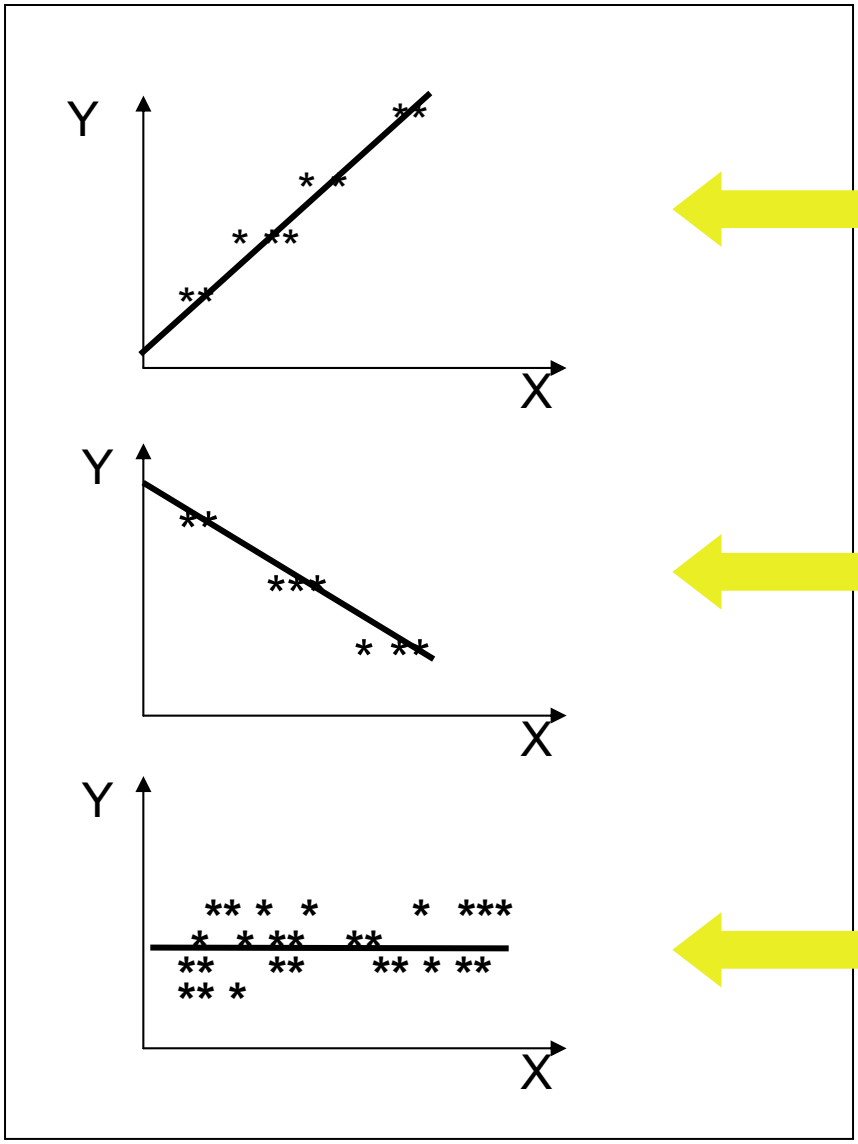
Régression linéaire
Modèle linéaire

Traités indifféremment dans la suite de ce cours grâce à l'utilisation de la méthode des moindres carrés

Le Modèle linéaire simple



Sens de la pente α

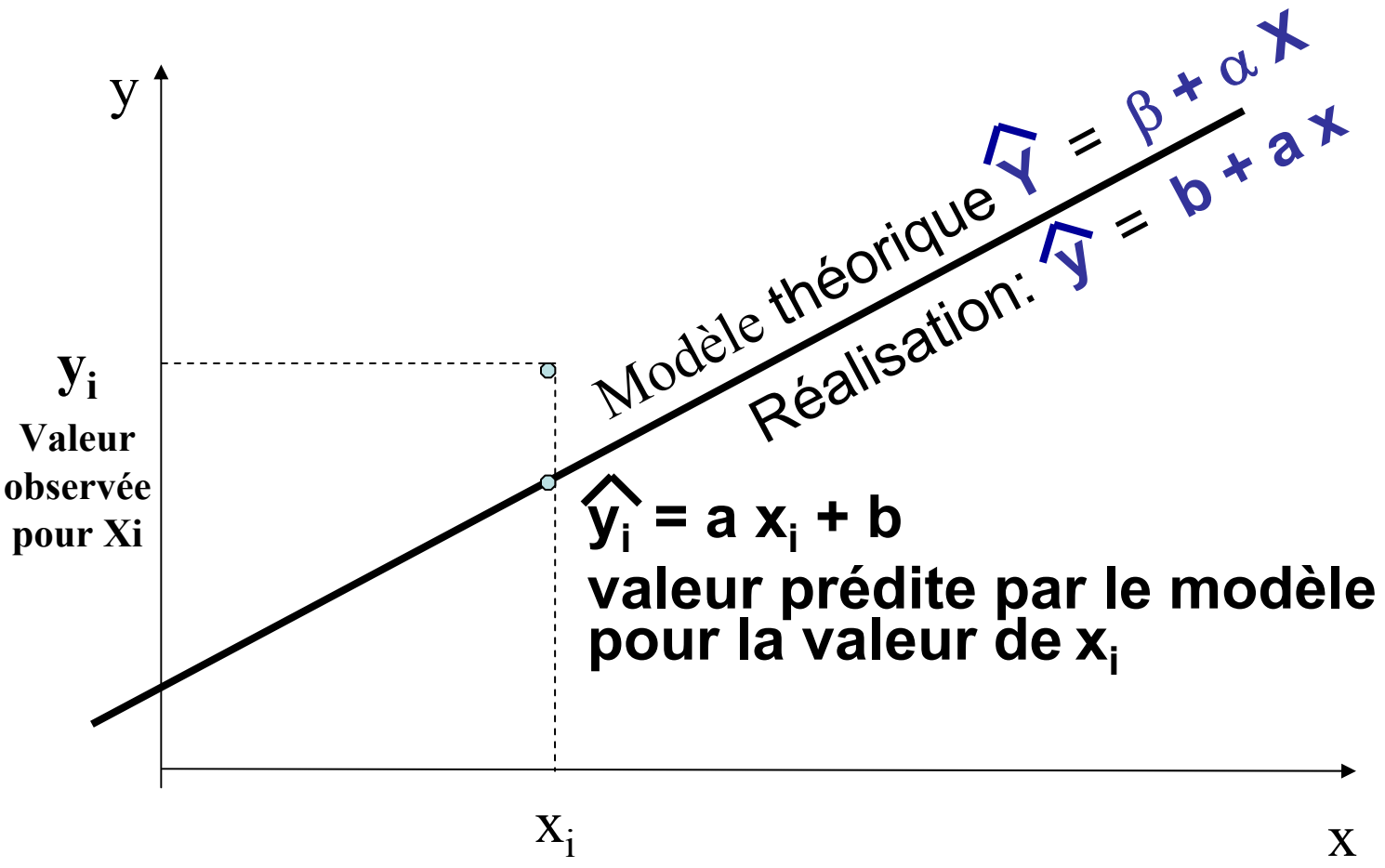


Relation positive entre X et Y:
Quand X augmente, Y augmente.
Un test devra être réalisé pour savoir si cette relation est significative

Relation négative entre X et Y:
Quand X augmente, Y diminue.
Un test devra être réalisé pour savoir si cette relation est significative

Pas de relation entre X et Y:
Les variations de Y ne dépendent pas des variations de X

Estimation des paramètres de la droite de régression

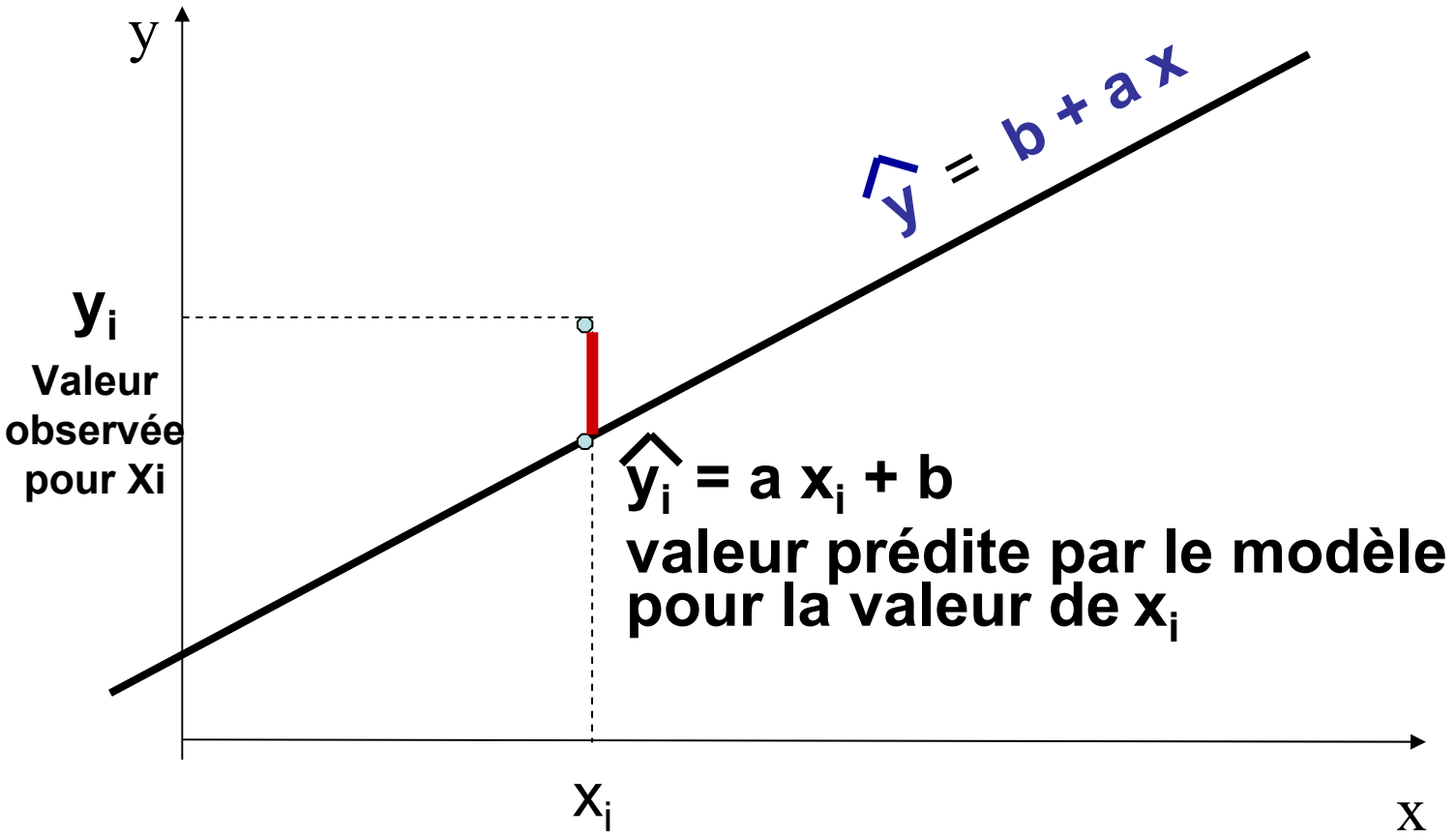


Au couple (x_i, y_i) observé s'ajoute $\hat{y}_i$ prédit par le modèle

Importance de l'erreur ε appelée aussi résidu, écart au modèle, erreur aléatoire.

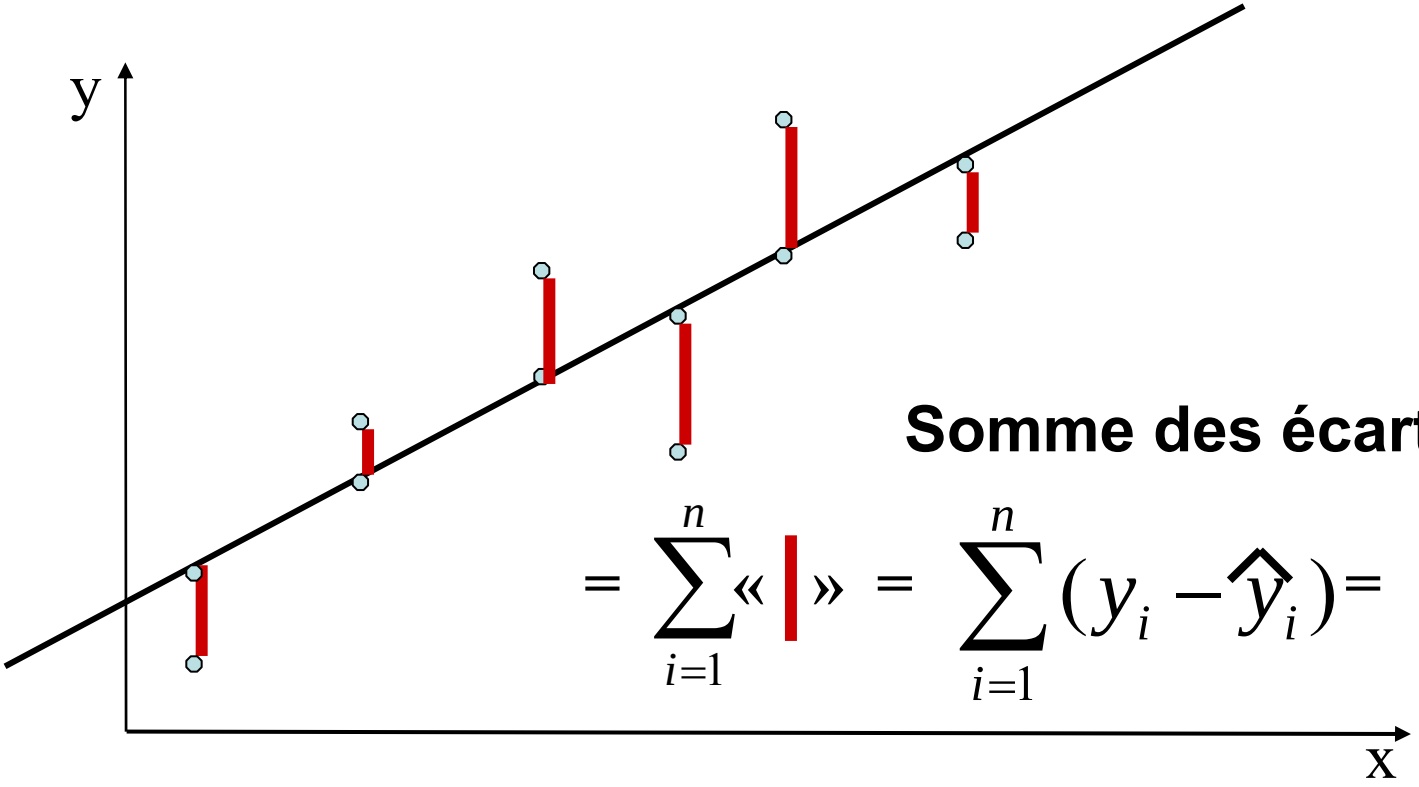
$$\begin{cases} Y = \hat{Y} + \varepsilon \\ \varepsilon = Y - \hat{Y} \\ e_i = y_i - \hat{y}_i \end{cases}$$

Estimation des paramètres de la droite de régression



L'écart « | » est égal à $y_i - \hat{y}_i$ soit égal à e_i

Estimation des paramètres de la droite de régression (prise en compte des n écarts)



Somme des écarts

$$= \sum_{i=1}^{n} \text{« | »} = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i$$

Somme des <u>carrés</u> des écarts (SCE) $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

Estimation des paramètres du modèle par le méthode des moindres carrés

Trouver pour le nuage des n points (x_i , y_i) une droite dont l'équation soit telle que SCE soit minimale

$$\mathbf{SCE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Deux inconnues: a et b

Connus: n couples (x_i ,y_i)

Méthode des moindres carrés

$$\text{SCE} = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \text{F(a ,b)}$$

Le minimum est atteint pour l'annulation des deux dérivées partielles:

$$\begin{cases} \dfrac{\partial SCE}{\partial a} = 0 \\ \dfrac{\partial SCE}{\partial b} = 0 \end{cases}$$

Méthode des moindres carrés

On obtient un système de deux équations à deux inconnues a et b

$$\begin{cases} \sum_{i=1}^{n} [2(y_i - (ax_i + b))(-x_i)] = 0 \\ \sum_{i=1}^{n} [2(y_i - (ax_i + b))(-1)] = 0 \end{cases}$$

Méthode des moindres carrés

Après simplification par - 2

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0 \end{cases}$$

Méthode des moindres carrés

On exprime b dans la seconde équation

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ b = \dfrac{\sum_{i=1}^{n} y_i}{n} - a\dfrac{\sum_{i=1}^{n} x_i}{n} \end{cases}$$

$$b = \bar{y} - a\bar{x}$$

Méthode des moindres carrés

On remplace b par sa valeur dans la 1er équation pour obtenir a et on × par $\frac{1}{n}$ numérateur et dénominateur

$$a = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{(\sum_{i=1}^{n} x_i)^2}{n^2}}$$

COVARIANCE (X,Y)

$$\mathrm{COV}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$
$$\mathrm{COV}(Y,X) = \mathrm{COV}(X,Y)$$
$$\mathrm{COV}(X,X) = \mathrm{VAR}(X)$$

$$\mathrm{VAR}(X+Y) = \mathrm{VAR}(X) + \mathrm{VAR}(Y) + 2\mathrm{COV}(XY)$$

$$\mathrm{COV}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}$$

La covariance s'exprime comme «unité de X × unité de Y»

Méthode des moindres carrés

On obtient finalement une solution unique pour la droite de régression dont le résultat général est:

$$\begin{cases} a = \dfrac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \\ b = \bar{Y} - a\bar{X} \end{cases}$$

La pente s'exprime comme

$$\frac{\text{unité de Y}}{\text{unité de X}}$$

L'ordonnée à l'origine s'exprime avec l'unité de Y

Ce résultat montre bien que X et Y ne sont pas interchangeables dans la régression

Méthode des moindres carrés

- La droite de régression passe par le point $(\bar{x}, \bar{y})$ qui est le centre de gravité du nuage de points

→ Le point $(\bar{x}, \bar{y})$ est solution de l'équation (vérification)

- Les résidus ont une moyenne nulle (e réalisation de ε)

→ $$\frac{\sum_{i=1}^{n} e_i}{n} = 0$$

Conditions d'application pour X et Y

- **Quantitatives** ? **examen des données**
- **Appariées** ? **examen des données (**attention aux données manquantes)

- **distribution normale du couple (X,Y)**
  soit binormalité de X , Y
  soit pour tout X, les Y ont une distribution normale et
  pour tout Y, les X ont une distribution normale

?

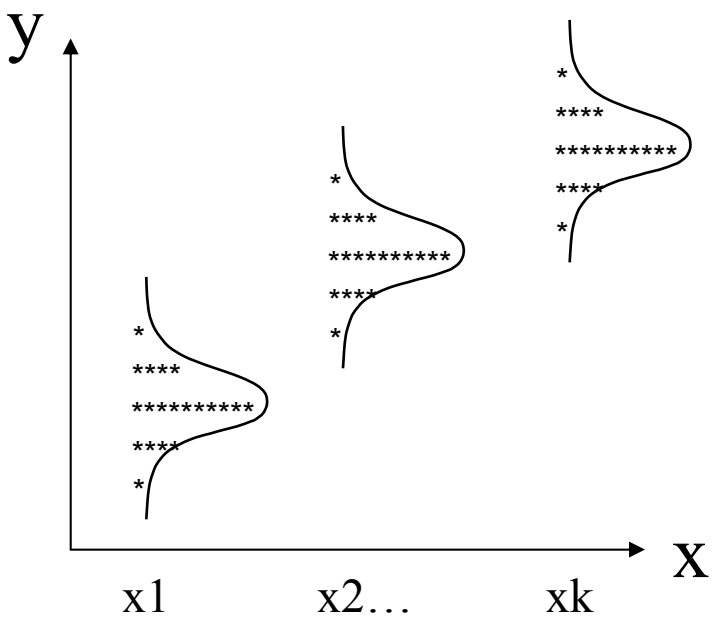
Rq: Si la binormalité est difficile à vérifier (souvent le cas):
distribution normale des X et distribution normale des Y

Conditions d'application pour X et Y

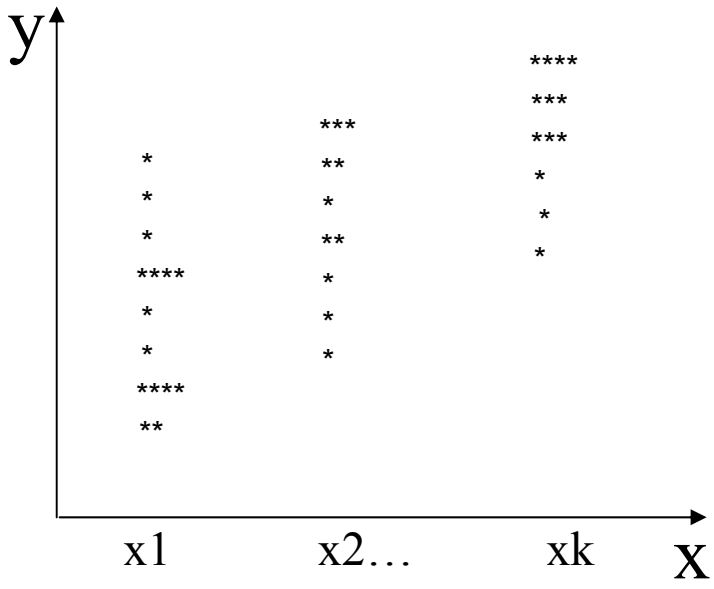
Ex: Pour chaque X, les Y sont-ils distribués normalement?

?

Etude des distributions des Y pour les valeurs de X (si répétitions de y pour un x).



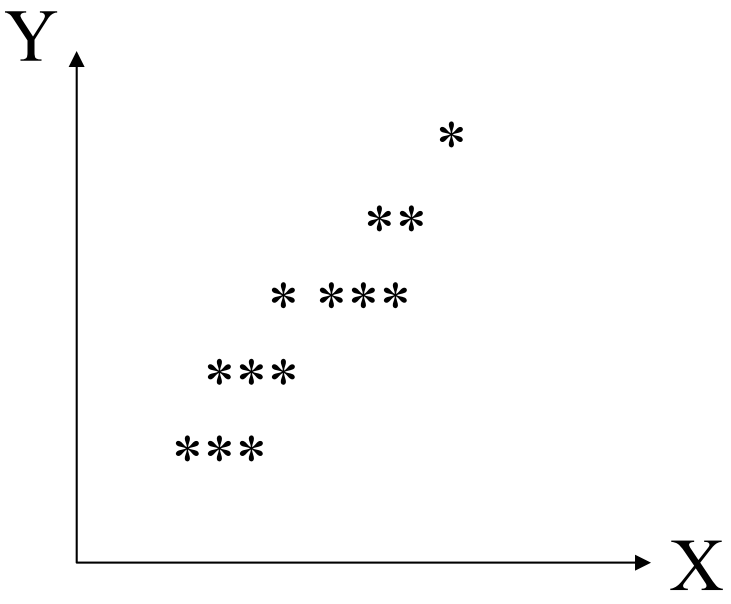
Distributions normales



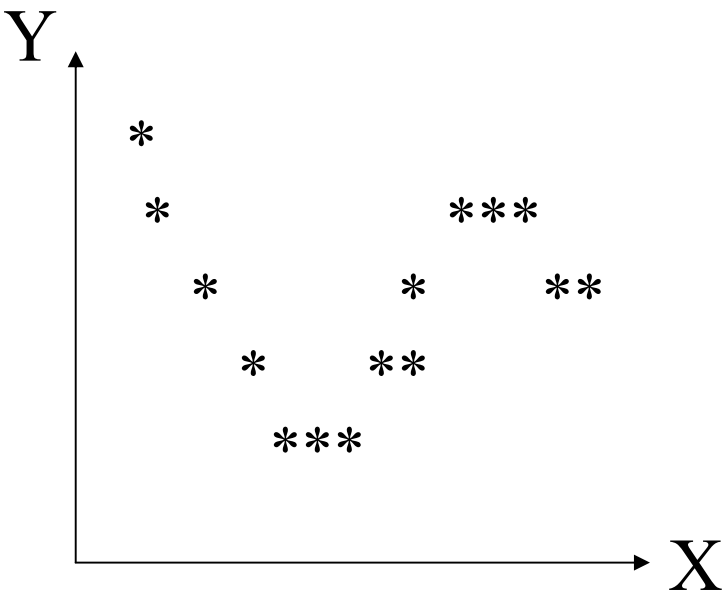
Distributions non normales

Contrôle de l'hypothèse de linéarité

- **Linéarité entre X et Y** ? Visualisation du nuage de points.



Linéarité plausible



Pas de linéarité

Contrôle des hypothèses dont dépendent les propriétés de la méthode des moindres carrés

Etude des résidus ε

Les résidus doivent :

1/avoir une distribution normale N (0 , σ)

2/avoir une variance constante quelle que soit la valeur de x (homoscédasticité)

3/ être indépendants (non autocorrélés). On ne peut pas déduire la valeur d'un résidu à partir des autres résidus (tests d'indépendance des résidus).

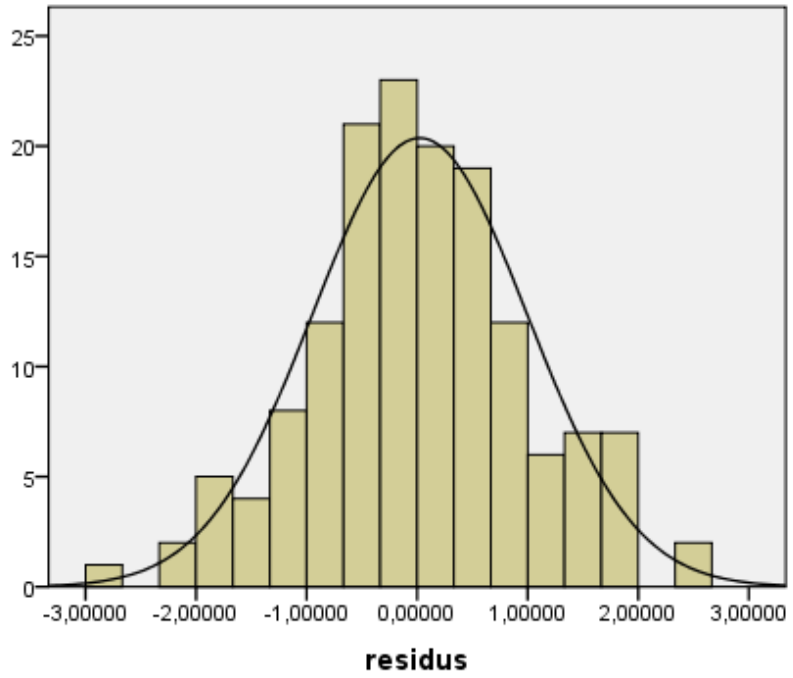
?

Vérification (souvent) empirique

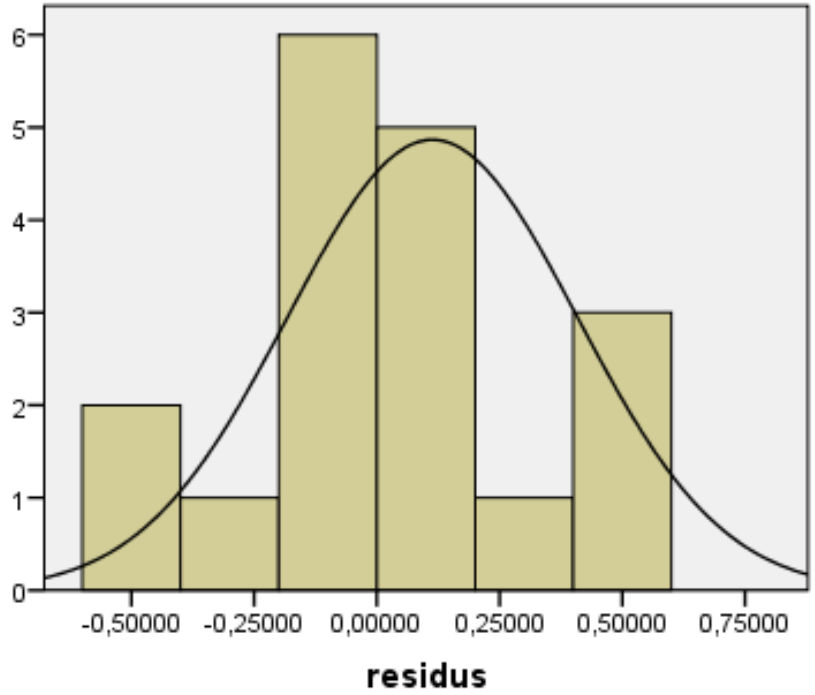
Contrôle des hypothèses dont dépendent les propriétés de la méthode des moindres carrés

Etude des résidus:

Normalité: Histogramme des résidus (aspect gaussien)



Normalité plausible

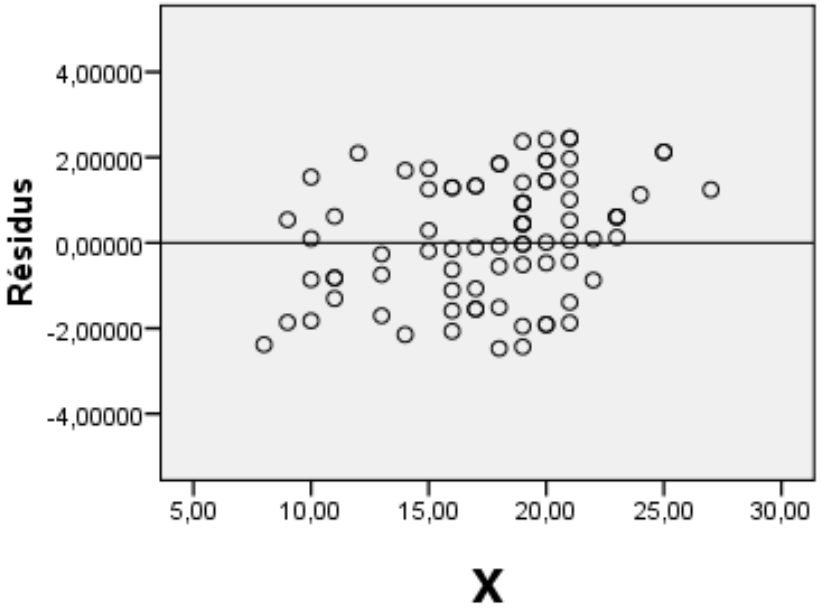


Pas de normalité

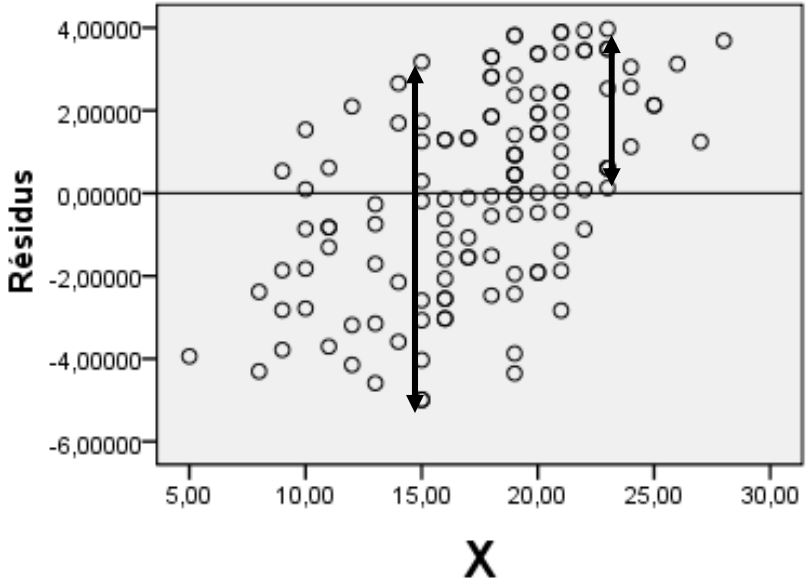
Contrôle des hypothèses dont dépendent les propriétés de la méthode des moindres carrés

Etude des résidus : Homoscédasticité

Graphe des résidus en fonction du prédicteur
(il ne doit pas apparaître de tendance)



Homoscédasticité plausible



Hétéroscédasticité

Coefficient de corrélation linéaire

X et Y ont des rôles interchangeables

$$\rho_{X,Y} = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{VAR}(X) \times \mathrm{VAR}(Y)}}$$

$$-1 \leq \rho_{X,Y} \leq 1$$

Coefficient de corrélation linéaire (réalisation)

$$-1 \leq r_{x,y} \leq 1$$

$$r_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}}{\sqrt{\left[\frac{\left(\sum_{i=1}^{n} x_i^{\,2}\right)}{n} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}\right] \times \left[\frac{\left(\sum_{i=1}^{n} y_i^{\,2}\right)}{n} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n^2}\right]}}$$

Mesure l'intensité de la liaison entre X et Y

$|r_{x,y}|$ Proche de 1 $\rightarrow \exists$ RELATION entre les variables

Estimation des variances des coefficients a et b

Variance résiduelle de la régression

$$\widehat{\sigma^2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Variance de a

$$\widehat{\sigma^2}_a = \frac{\widehat{\sigma^2}}{ns_x^2}$$

Test de la pente à 0

$H_0 : \alpha = 0$

H1: $\alpha \neq 0$

(attention α correspond ici à la pente et non au risque du même nom)

$$t_{(n-2)ddl} = \frac{a}{\hat{\sigma}_a}$$

soit

$$t_{(n-2)ddl} = \frac{a}{\sqrt{\frac{\hat{\sigma}^2}{ns_x^2}}}$$

Pour le risque de première espèce donné (en général égal à 5%)

Si la valeur calculée du test de Student < valeur tabulée pour (n-2) ddl : rejet de H1

Si la valeur calculée du test de Student ≥ valeur tabulée pour (n-2) ddl : rejet de H0

Test de la pente à une valeur α_1

$H_0 : \alpha = \alpha 1$

H1: $\alpha \neq \alpha 1$

(attention α correspond ici à la pente et non au risque du même nom)

$$t_{(n-2)ddl} = \frac{|a - \alpha_1|}{\hat{\sigma}_a}$$

soit

$$t_{(n-2)ddl} = \frac{|a - \alpha_1|}{\sqrt{\frac{\hat{\sigma}^2}{ns_x^2}}}$$

Pour le risque de première espèce donné (en général égal à 5%)

Si la valeur calculée du test de Student < valeur tabulée pour (n-2) ddl : rejet de H1

Si la valeur calculée du test de Student ≥ valeur tabulée pour (n-2) ddl : rejet de H0

Test du coefficient de corrélation à 0

$H_0 : \rho = 0$
H1: $\rho \neq 0$

Estimation de la variance de r

$$\widehat{\sigma_r}^2 = \frac{1-r^2}{n-2}$$

$$t_{(n-2)ddl} = \frac{|r-\rho|}{\widehat{\sigma}_r}$$

soit

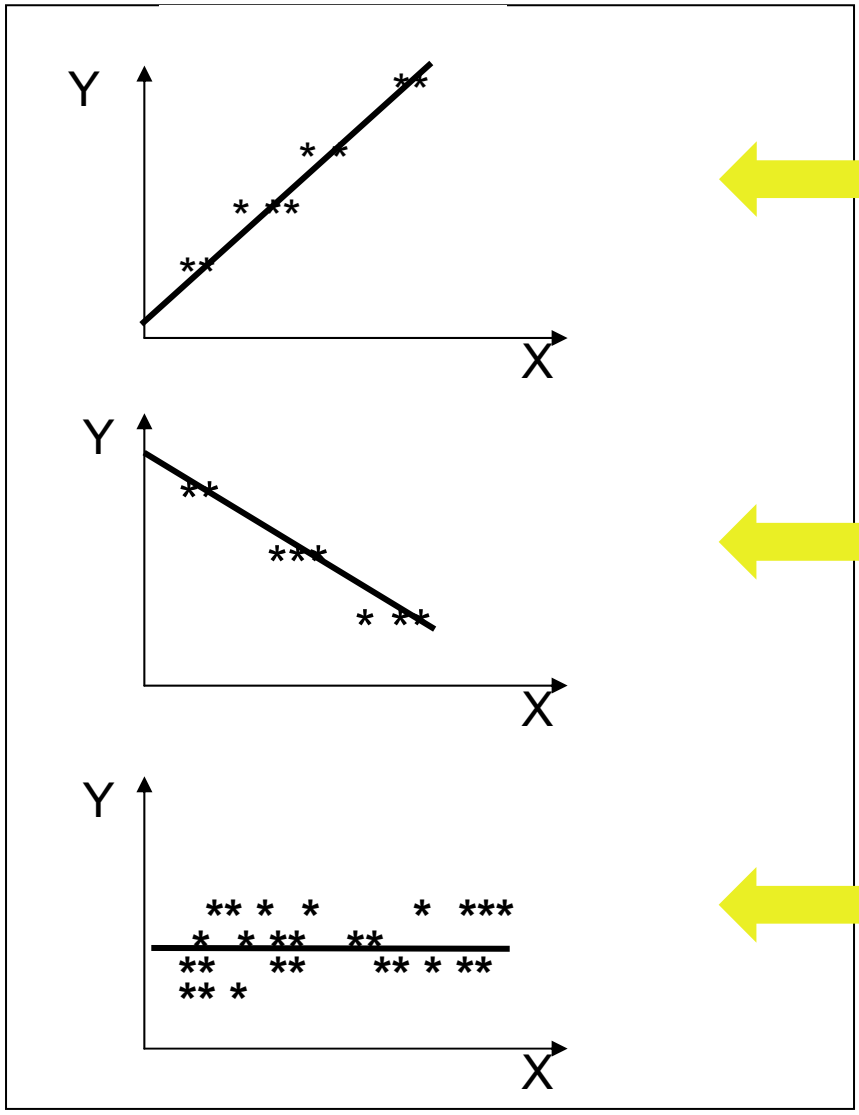
$$t_{(n-2)ddl} = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$

Pour le risque de première espèce donné (en général égal à 5%)
Si la valeur calculée du test de Student < valeur tabulée pour (n-2) ddl : rejet de H1
Si la valeur calculée du test de Student ≥ valeur tabulée pour (n-2) ddl : rejet de H0

Remarque sur les tests

Le test de la pente à 0 et
le test du coefficient de corrélation à 0
donnent
la même conclusion

Sens de la pente α ou du coefficient r



Covariance positive Pente positive
Coefficient de corrélation positif
Relation significative si le test de la pente à 0 (ou le test du coefficient de corrélation à 0) est significatif

Covariance négative, Pente négative
Coefficient de corrélation négatif
Relation significative si le test de la pente à 0 (ou le test du coefficient de corrélation à 0) est significatif

Pas de relation entre X et Y:
Les variations de Y ne dépendent pas des variations de X
Le test de la pente à 0 et le test du coefficient de corrélation à 0 sont non significatifs

Relation a a' r (aspect théorique)

Sur les n mêmes couples (x,y)

$$\hat{y} = ax + b \quad \begin{cases} a = \dfrac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \\ b = \bar{Y} - a\bar{X} \end{cases}$$

$$\hat{x} = a'y + b' \quad \begin{cases} a' = \dfrac{\mathrm{cov}(Y,X)}{\mathrm{var}(Y)} \\ b' = \bar{X} - a'\bar{Y} \end{cases}$$

$$r_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X) \times \mathrm{var}(Y)}}$$

- Même signe pour a, a' et r (celui de cov(x,y))
- Les deux droites se coupent au point $(\bar{x}, \bar{y})$
- La valeur absolue du coefficient de corrélation linéaire est égal à la moyenne géométrique des pentes.

$$|r| = \sqrt{aa'}$$

Prédiction de Y à partir de l'équation de la droite : Interpolation, extrapolation

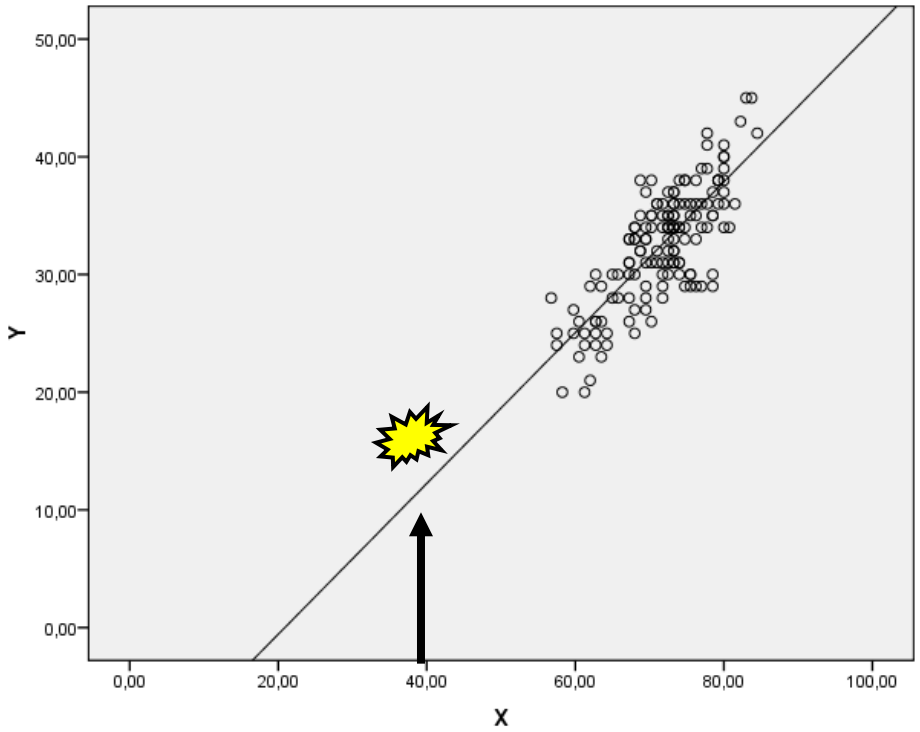
Une fois connue l'équation de la droite :

$$\hat{y} = b + a\,x$$

il est possible de calculer la valeur de $\hat{y}$ pour x donné

Attention :

Modèle linéaire plausible.
Rester dans les limites +++



Exemple

- Position du problème: En préalable à une intervention chirurgicale, la mesure de L1 (longueur en mm) et la mesure de L2 (angle en degrés) sont effectuées à partir de résultats d'imagerie. La mesure L2 est particulièrement importante pour préparer l'intervention mais son obtention est plus difficile que celle de la mesure L1.

- Objectif: Le but est de savoir si la mesure L1 pourrait aider à prédire la mesure L2

Exemple

- Les mesures L1 et L2 ont été réalisées sur n=149 sujets. On admettra que la binormalité (L1,L2) est respectée.
- Les valeurs suivantes sont obtenues:

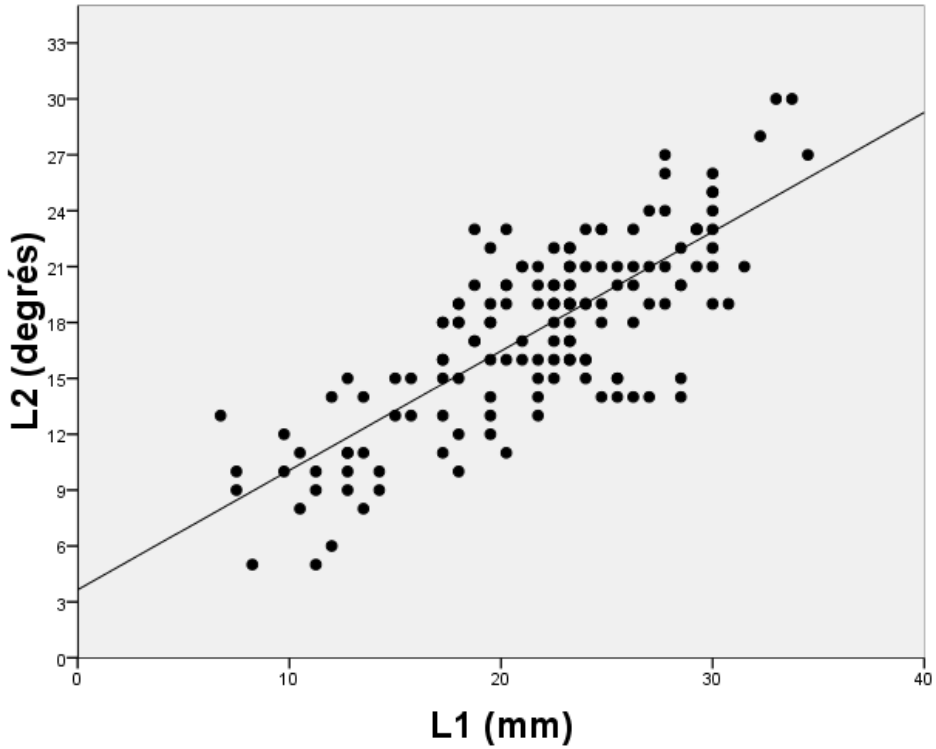
$$m_{L1} = 21.65mm \qquad s_{L1}^{\ 2} = 36.46mm^2$$

$$m_{L2} = 17.52\,\text{degrés} \qquad s_{L2}^{\ 2} = 24.48\,\text{degrés}^2$$

$$\text{cov}_{L1,L2} = 23.35\text{mm} \times \text{degrés}$$

$$\widehat{\sigma_a^2} = 0.00177\text{degrés}^2/\text{mm}^2$$

Régression L2=aL1+b

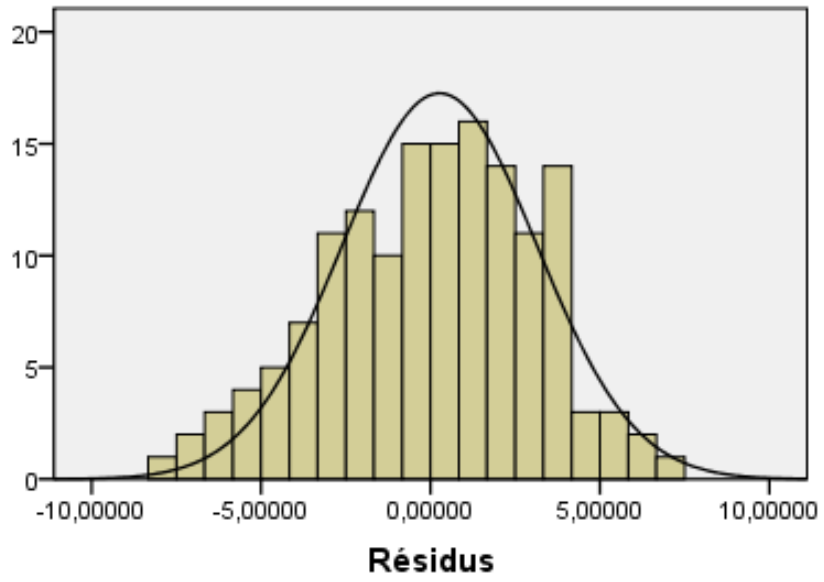


a = 0.640 degrés/mm

b = 3.655 degrés

Equation de la droite: L2(degrés) = 0.640 L1 + 3.655

Résidus et Tests (α=5%)



Normalité plausible

- Test de la pente à 0

$t_{\alpha,(147)ddl}$ = 15.2 (>1.96 Significatif)

- Prédiction de L2 pour L1=15mm

13,26 degrés

- Prédiction de L2 pour L1=50mm

Hors limites

- r = 0.782
- Test de r à 0

$t_{\alpha,(147)ddl}$ = 15.2 (>1.96 Significatif)

Régression L2=aL1+b

- Les tests (pente à 0 et r à 0) montrent qu'il existe une relation significative entre L1 et L2, cette relation étant positive (signe de la covariance).
- La valeur de L1 peut ainsi aider à prédire la valeur de L2.
- Remarque: Attention, cela ne prouve pas pour autant que l'estimation de L2 obtenue à partir de la mesure de L1 puisse *remplacer* la mesure de L2. Des tests non traités dans ce cours sont alors nécessaires pour aller plus loin dans cette démarche particulière.

Nancy-Université
Université Henri Poincaré

ministère éducation nationale enseignement supérieur recherche